

Manual de Usuario del Programa para Tokenización de Texto



TI-1400 Introducción a la Programación

Segundo Proyecto Programado

I Semestre, 2020

Integrantes:

Jose Altamirano Salazar - 2020426159

Josué Brenes Alfaro -2020054427

Introducción:

El Programa para Tokenización de Texto, se presenta como una herramienta para iniciar el proceso de análisis de un texto. Esto se realiza a través de la clasificación de los componentes de las oraciones y su presentación estructurada. Partiendo de un proceso conocido como tokenización, que se encargará de categorizar cada una de las palabras según sea un artículo, preposición, pronombre, verbo número o que no pertenezca a ninguna de las anteriores.

Además, cuenta con la funcionalidad para ingresar texto de diferentes maneras. El usuario podrá ingresarlo de forma manual o a través de la lectura de archivos de texto que se encuentren en su computador.

Por otra parte, el programa brinda al usuario la capacidad de traducir algunos de los elementos que compongan el texto. Esta operación le ofrece una comprensión mucho más amplia sobre el contenido del documento, al proporcionarle un entendimiento de algunas de las partes del texto incluso dentro del idioma inglés. Esta traducción puede ser consultada desde el programa o desde los archivos generados por el este.

Adicional a las funciones mencionadas, el programa cuenta con la utilidad de generar un archivo HTML que contiene la estructuración del texto generada dentro del programa. Esta función sobre la portabilidad permite al usuario almacenar los tokens extraídos para poder consultarlos posteriormente. La flexibilidad que ofrecen los HTML al poder interpretarse dentro de cualquier sistema operativo, al contar solamente con un navegador web, le permite al usuario acceder a los archivos generados cuando y donde lo necesite.

La ventana principal tiene un enfoque visual atractivo al usuario, con una funcionalidad intuitiva, para que cualquier persona pueda usarlo. La paleta de colores empleada para el diseño utiliza una combinación que no se convierte en un inconveniente a la vista para navegar dentro del programa. La interacción con el programa se realiza por medio de un menú de opciones en la parte superior y una serie botones, los cuales indican su propósito en la leyenda que contienen.

Funcionalidad General:

La funcionalidad principal del programa de tokenización se centra en realizar la clasificación de un texto, según el origen morfológico de las palabras que integran cada oración recibida. Con esto se hace referencia a la categorización derivado de realizar un análisis morfológico de un texto. En el caso del programa esta labor la realiza por medio de la separación de cada elemento de del texto recibido en partículas únicas denominadas tokens, a través de un proceso de comparación de cada una de estas con sus posibles valores dentro de las categorías permitidas.

En este caso, las categorías mencionadas corresponden a las siguientes:

- Artículos: el, la, los, las, un, una, unos, unas, lo, del y al.
- Preposiciones: a, ante, bajo, cabe, con, contra, de, desde, durante, en, entre, hacia, hasta, mediante, para, por, según, sin, so, sobre, tras, versus, vía
- Pronombres yo, me, mí, conmigo, nosotros, nosotras, nos, tú, te, ti, contigo, vosotros, vosotras, vos, él, ella, se, consigo, le, les, mío, mía, míos, mías, nuestro, nuestra, nuestros, nuestras, tuyo, tuya, tuyos, vuestro, vuestra, vuestros, vuestras, suyo, suya, suyos, suyas.
- Formas Verbales:
 - Infinitivos: verbos terminados en ar, er, ir
 - Gerundio: verbos terminados en ando, iendo
 - Participio: verbos terminados en ado, ido, to, so, cho
- Números: en su representación numérica de base 10 (0, 1, 2, 3, ... 9)
- Sin clasificador: cualquier otro token no identificado en las categorías anteriores.

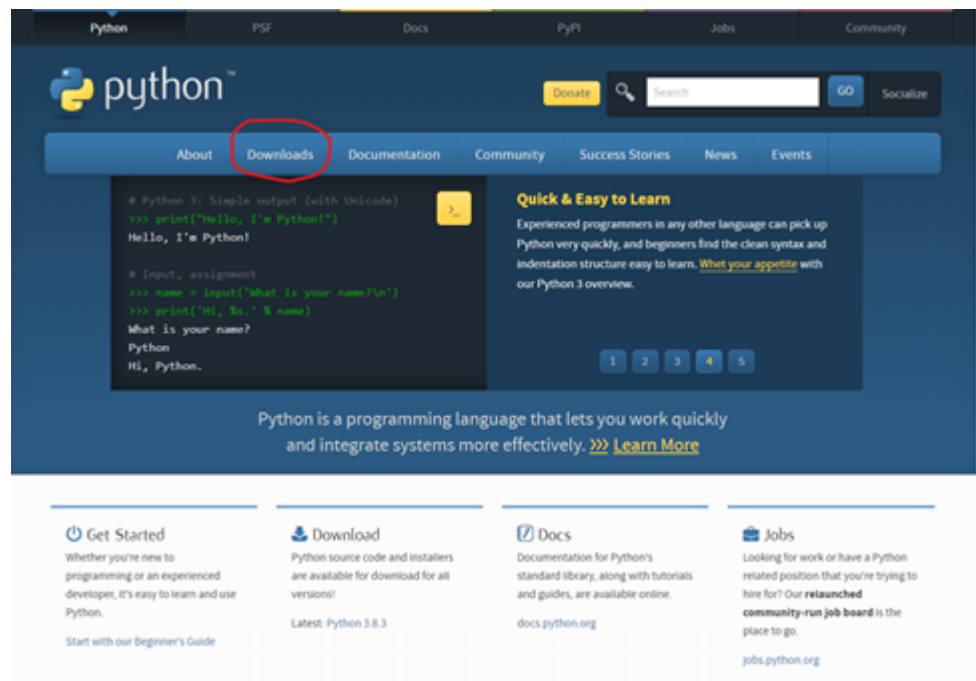
De esta manera, en caso de que alguna de las palabras extraídas del texto ingresado cumpla con las características propias de alguna de las categorías anteriores, esta será incluida en una lista y posteriormente desplegada para que pueda ser visualizada por el usuario del programa.

Otra función del programa corresponde la clasificación y traducción de algunos de los tokens del texto recibido. En este caso se realiza la traducción de los que correspondan a las categorías de artículos, pronombres preposiciones y verbos.

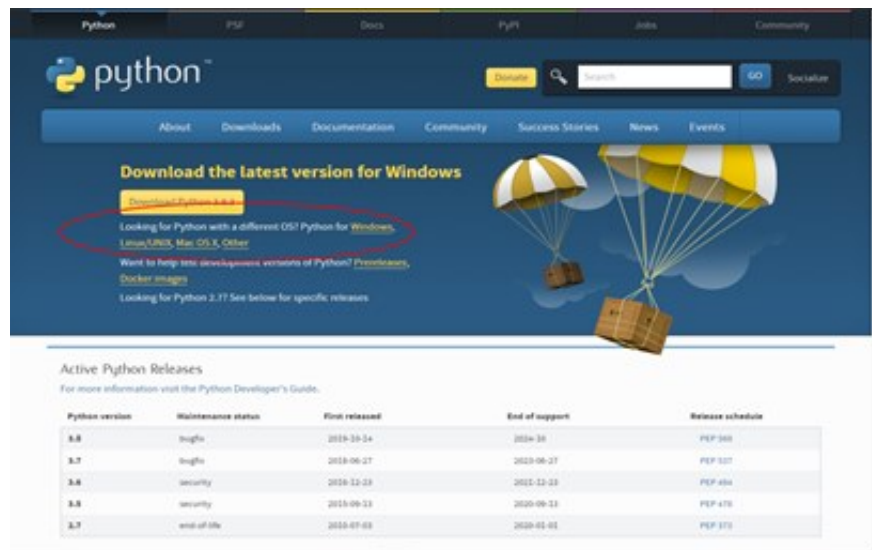
Finalmente, el programa también dispone de la posibilidad de almacenar la clasificación generada dentro de un archivo de tipo HTML. Dentro de este archivo se incluye una tabulación de cada una de las categorías previamente mencionadas para que pueda ser transportada o consultada sin necesidad de volver a procesar un texto en específico.

Requisitos del Programa:

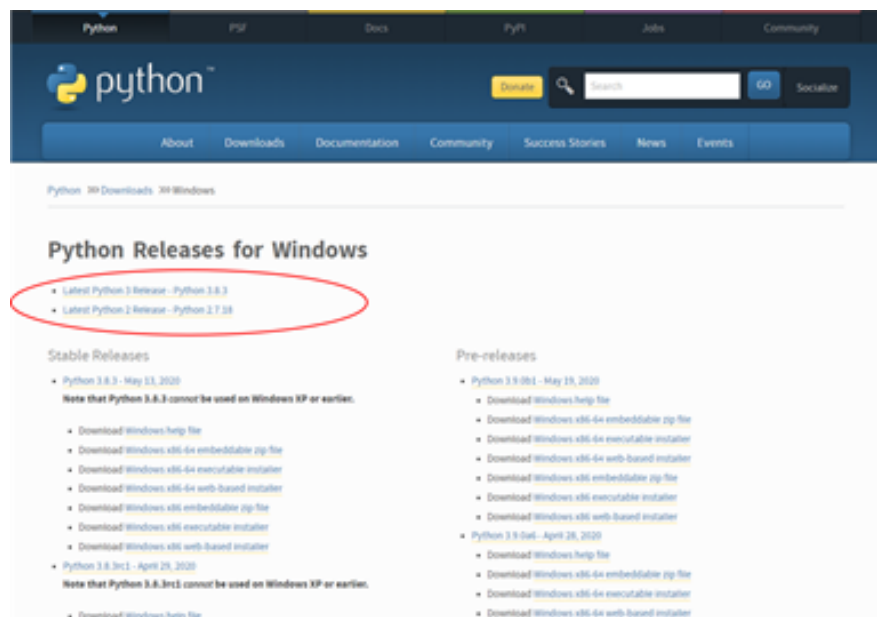
Para poder correr el programa debe tener un intérprete de Python. El programa se encuentra desarrollado bajo la versión 3.8, por lo cual es necesario descargar cualquier versión que iguale o supere a esta. Esta descarga se realiza por medio de la página oficial de Python (<https://www.python.org/>). Al ingresar a esta, es necesario navegar a la opción de descargas en la barra de menú superior.



Al ingresar a esta sección, es necesario que seleccione el sistema operativo de su equipo para realizar la descarga apropiada.



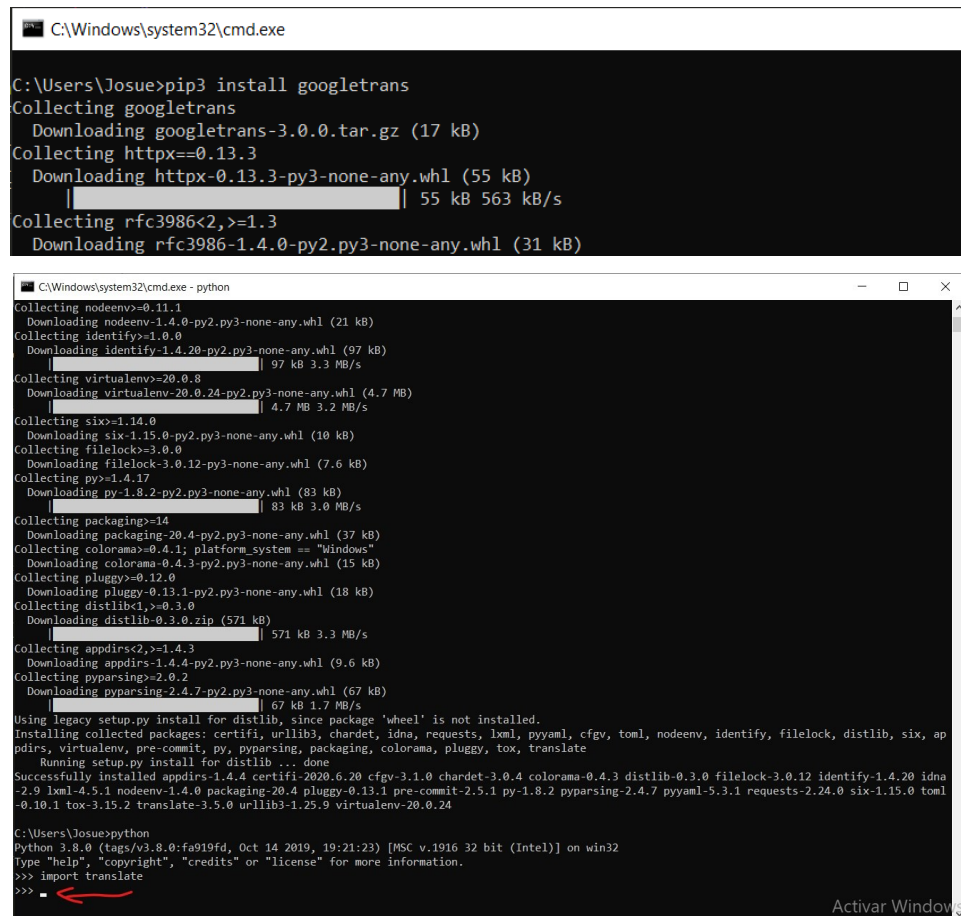
En cualquiera de los casos debe seleccionar la opción que coincida o supera la versión 3.8. Generalmente se incluye en la parte superior de las descargas de cada sistema operativo, tal como se muestra en la siguiente imagen para el caso de Windows.



Una vez que el ejecutable de Python se haya descargado proceda a instalarlo. Cuando esta instalación haya finalizado, será necesario instalar la librería para el correcto funcionamiento del programa. La primera es la librería “**Translator**”, para extraer la funcionalidad del API de traducción de google a través. Para instalar abra una ventana de la consola de Windows. Esto puede hacerlo desde el menú de inicio,

buscando el programa CMD o con la combinación de las teclas “Windows” y la tecla R, y luego escribiendo CMD en la ventana que se despliega.

Cuando se ejecute la consola de Windows, deberá ingresar el comando **pip3 install googletrans**, y presionar la tecla Enter. luego de esto se iniciará a instalación de la librería. Al finalizar se desplegará la línea para ingresar un nuevo comando en caso de que no haya sucedido ningún inconveniente.



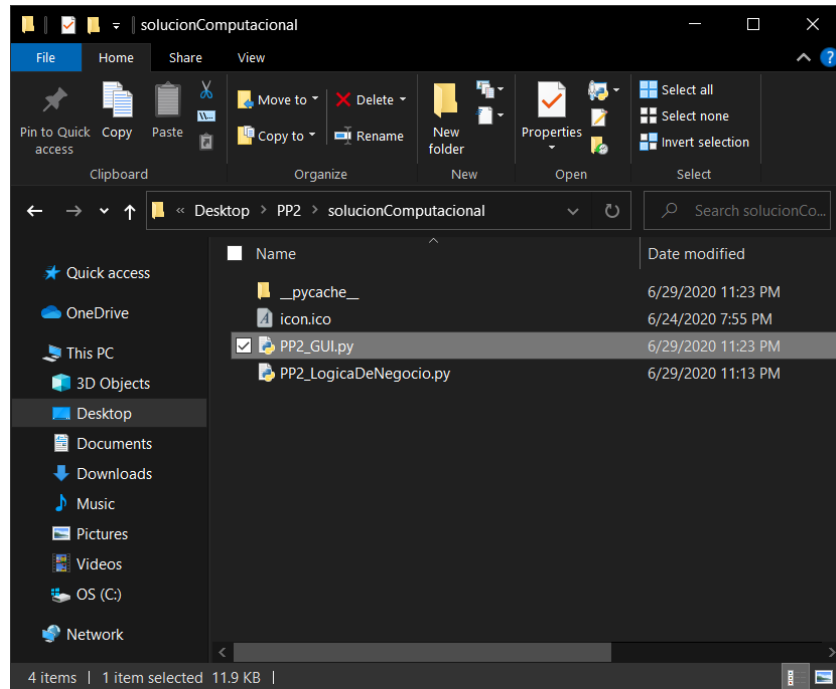
```
C:\Windows\system32\cmd.exe

C:\Users\Josue>pip3 install googletrans
Collecting googletrans
  Downloading googletrans-3.0.0.tar.gz (17 kB)
Collecting httpx==0.13.3
  Downloading httpx-0.13.3-py3-none-any.whl (55 kB)
    | 55 kB 563 kB/s
Collecting rfc3986<2,>=1.3
  Downloading rfc3986-1.4.0-py2.py3-none-any.whl (31 kB)

C:\Windows\system32\cmd.exe - python
Collecting nodeenv==0.11.1
  Downloading nodeenv-1.4.0-py2.py3-none-any.whl (21 kB)
Collecting identify==1.0.0
  Downloading identify-1.4.20-py2.py3-none-any.whl (97 kB)
    | 97 kB 3.3 MB/s
Collecting virtualenv==20.0.8
  Downloading virtualenv-20.0.24-py2.py3-none-any.whl (4.7 MB)
    | 4.7 MB 3.2 MB/s
Collecting six==1.14.0
  Downloading six-1.15.0-py2.py3-none-any.whl (10 kB)
Collecting filelock==3.0.0
  Downloading filelock-3.0.12-py3-none-any.whl (7.6 kB)
Collecting py==1.4.17
  Downloading py-1.8.2-py2.py3-none-any.whl (83 kB)
    | 83 kB 3.0 MB/s
Collecting packaging==14
  Downloading packaging-20.4-py2.py3-none-any.whl (37 kB)
Collecting colorama==0.4.1; platform_system == "Windows"
  Downloading colorama-0.4.3-py2.py3-none-any.whl (15 kB)
Collecting pluggy==0.12.0
  Downloading pluggy-0.13.1-py2.py3-none-any.whl (18 kB)
Collecting distlib<1,>=0.3.0
  Downloading distlib-0.3.0.zip (571 kB)
    | 571 kB 3.3 MB/s
Collecting appdirs<2,>=1.4.3
  Downloading appdirs-1.4.4-py2.py3-none-any.whl (9.6 kB)
Collecting pyparsing==2.0.2
  Downloading pyparsing-2.4.7-py2.py3-none-any.whl (67 kB)
    | 67 kB 1.7 MB/s
Using legacy setup.py install for distlib, since package 'wheel' is not installed.
Installing collected packages: certifi, urllib3, chardet, idna, requests, lxml, pyyaml, cfgv, toml, nodeenv, identify, filelock, distlib, six, appdirs, virtualenv, pre-commit, py, pyparsing, packaging, colorama, pluggy, tox, translate
  Running setup.py install for distlib ... done
Successfully installed appdirs-1.4.4 certifi-2020.6.20 cfgv-3.1.0 chardet-3.0.4 colorama-0.4.3 distlib-0.3.0 filelock-3.0.12 identify-1.4.20 idna-2.9 lxml-4.5.1 nodeenv-1.4.0 packaging-20.4 pluggy-0.13.1 pre-commit-2.5.1 py-1.8.2 pyparsing-2.4.7 pyyaml-5.3.1 requests-2.24.0 six-1.15.0 toml-0.10.1 tox-3.15.2 translate-3.5.0 urllib3-1.25.9 virtualenv-20.0.24

C:\Users\Josue>python
Python 3.8.0 (tags/v3.8.0:fa919fd, Oct 14 2019, 19:21:23) [MSC v.1916 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> import translate
>>>
```

Ahora diríjase a la ubicación donde haya almacenado el archivo ejecutable del programa de Tokenización. Al hacer doble clic sobre el archivo **“PP2_GUI.py”**, se iniciará su ejecución. La siguiente imagen muestra este resultado dentro del SO Windows.

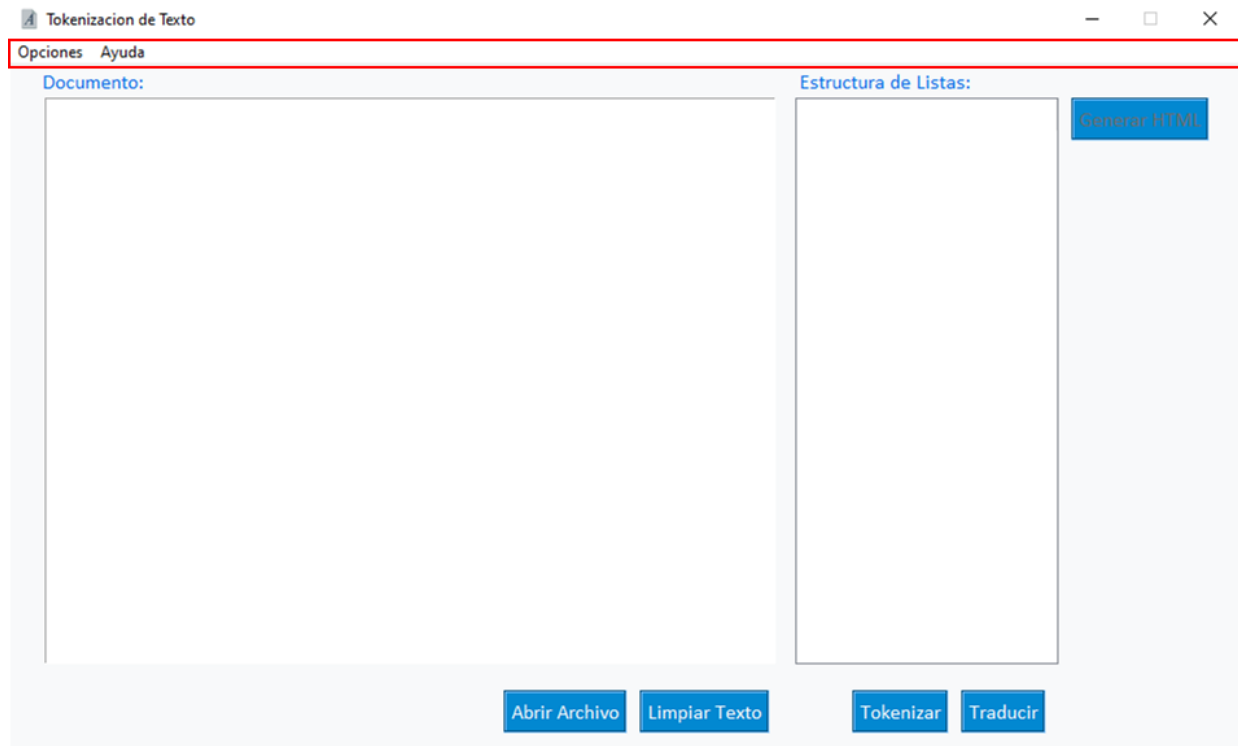


Guía de Uso:

El programa de Tokenización se compone de una única ventana principal y de algunos mensajes emergentes para indicar al usuario sobre el estado de los procesos o tareas que se estén desempeñando. Para una mejor comprensión sobre las secciones de la ventana y de las opciones de las cuales dispone el programa se ofrece el siguiente desglose explicativo.

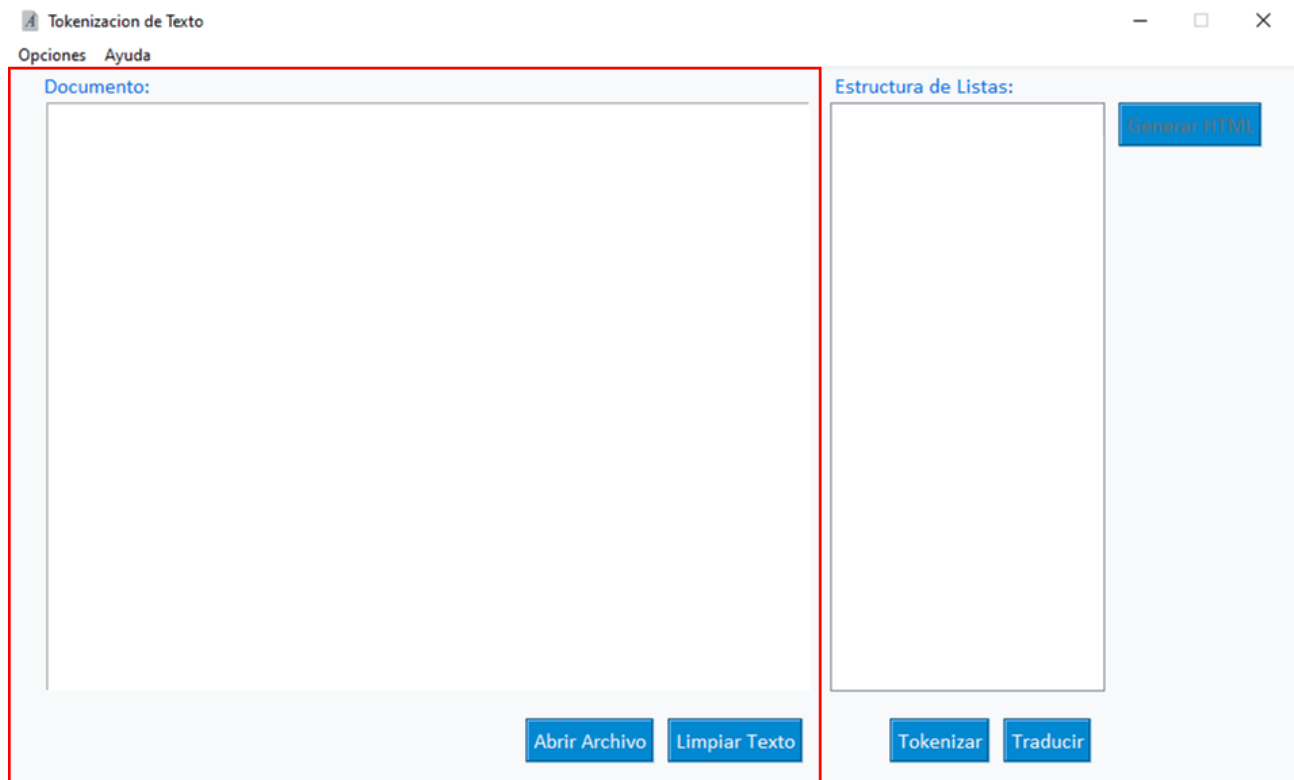
- **Secciones del Programa**

La ventana principal del programa puede dividirse en tres secciones principales. La primera corresponde a la barra de menú. Esta se encuentra en la parte superior de la pantalla, permite acceder a algunas funciones del programa, solicitar una guía sobre el uso de este, o cerrarlo. Su funcionalidad y secciones se explicarán de una forma más amplia en el siguiente apartado.

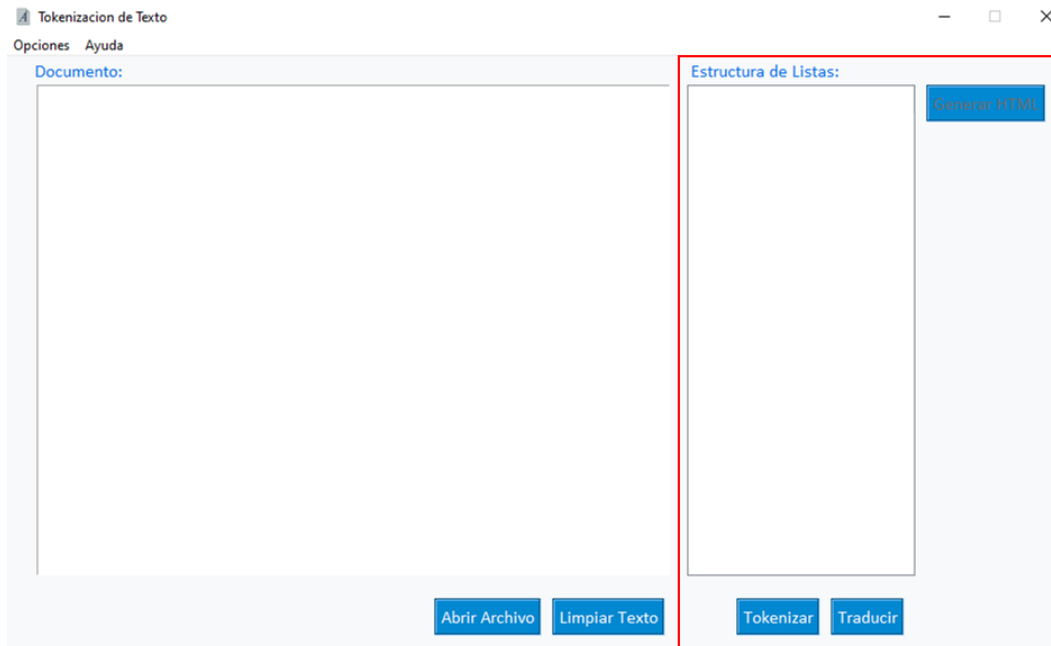


La segunda sección corresponde a la de “**Documento**”, en esta se realiza ingreso y la edición de texto. Se ubica en el lado izquierdo de la ventana principal y contiene principalmente un campo de texto. En este es posible ingresar texto de forma manual desde el teclado del computador o copiando este desde otra ventana. Es también en este espacio donde se mostrará el texto que está contenido en los archivos de texto que sean requeridos y abiertos dentro del programa.

Aun dentro de la sección de “**Documento**”, en la parte inferior se encuentran dos botones. El primero, con la leyenda “**Abrir Archivo**”. Al dar clic sobre este desplegará el explorador de archivos desde el cual podrá seleccionarse un archivo de texto para que su contenido sea visualizado dentro del programa. Junto al botón anterior se encuentra el de “**Limpiar texto**”. Al presionar este botón se borrará el contenido del campo de texto.



La siguiente sección corresponde a la de “**Estructura de Listas**” y se ubica en al lado derecho de la sección anterior. Esta se compone de un espacio para desplegar las listas de tokens que el programa vaya generando, dos botones para interactuar con el texto ingresado o generar el listado que será desplegado, y otro botón para generar el archivo HTML, con las listas de tokens tabuladas. Estos botones corresponden a lo que contienen las leyendas “**Tokenizar**” y “**Traducir**” en la parte inferior del campo de listado, y al botón “**Generar HTML**” al costado derecho de este.



El botón de “**Tokenizar**” se encarga de realizar la clasificación correspondiente sobre cada uno de los elementos del texto que haya sido ingresado. Luego de esto, las listas con los artículos, pronombres, preposiciones, verbos, números y demás palabras serán desplegadas en el campo de listado de la sección izquierda de la pantalla.

El botón de “**Traducir**” realiza una clasificación similar al botón anterior, pero en este caso los tokens serán desplegados en inglés y las categorías a utilizar son solamente las listas con los artículos, pronombres, preposiciones y verbos. De igual manera estas palabras serán desplegadas en el campo de listado de la sección izquierda de la pantalla.

Finalmente, el botón de “**Generar HTML**” se encarga de iniciar el proceso de construcción y guardado del archivo HTML que contiene una clasificación tabulada de las palabras desplegadas en los listados de tokens de la sección izquierda de la pantalla. Este archivo contiene el texto ingresado dentro del programa, seguido de una tabla con todas las clasificaciones y en sus columnas las palabras que pertenecen a cada una de estas. Debajo de esta tabla se incluye también otra con las clasificaciones válidas para los tokens que puedan ser traducidos a inglés, los cuales, tal como se indicó para el botón “**Traducir**”, serían solamente cuatro categorías.

- **Menú de Opciones Superior**

En la parte superior de la ventana principal puede observarse una barra de menú con dos opciones principales: **“Opciones”** y **“Ayuda”**.

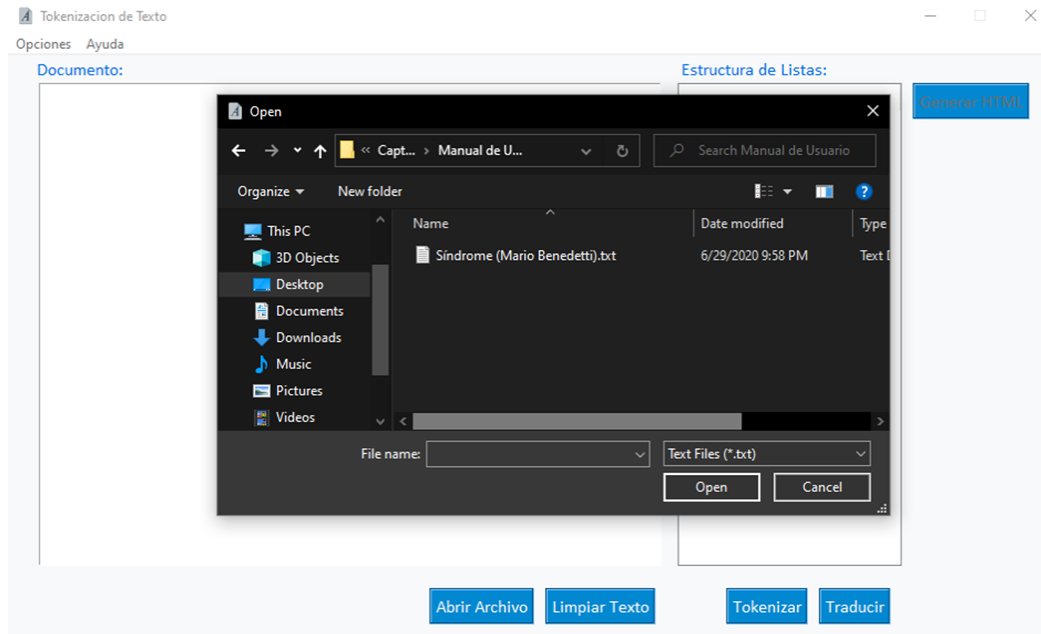


Cada una de las opciones mencionadas corresponde a su vez a un menú desplegable. Al hacer clic sobre la opción de **“Opciones”**, desplegarán algunas de las tareas que puede realizar el programa.



La primera de estas corresponde a **“Abrir un archivo”**, al hacer clic sobre esta se abrirá la ventana del explorador de archivos, desde la cual es posible seleccionar un archivo de texto o con la extensión “.txt” para que su contenido sea extraído por el programa.





Desde la segunda opción del primer menú desplegable, “**Borrar valores**”, es posible borrar los valores que hayan sido ingresados dentro del programa. en este caso estos valores pueden ser texto que el usuario haya ingresado, texto que se haya leído desde un archivo o incluso los resultados de traducir y tokenizar alguna de las opciones de texto mencionadas.



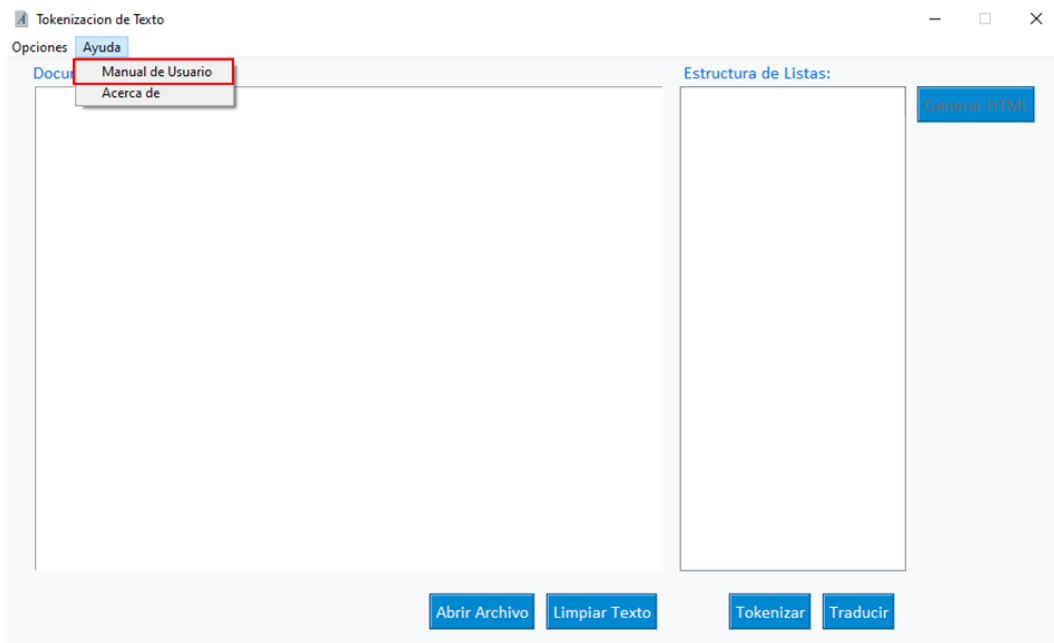
Desde la tercera opción del primer menú desplegable, “**Salir del programa**”, es posible cerrar el programa o detener su ejecución.

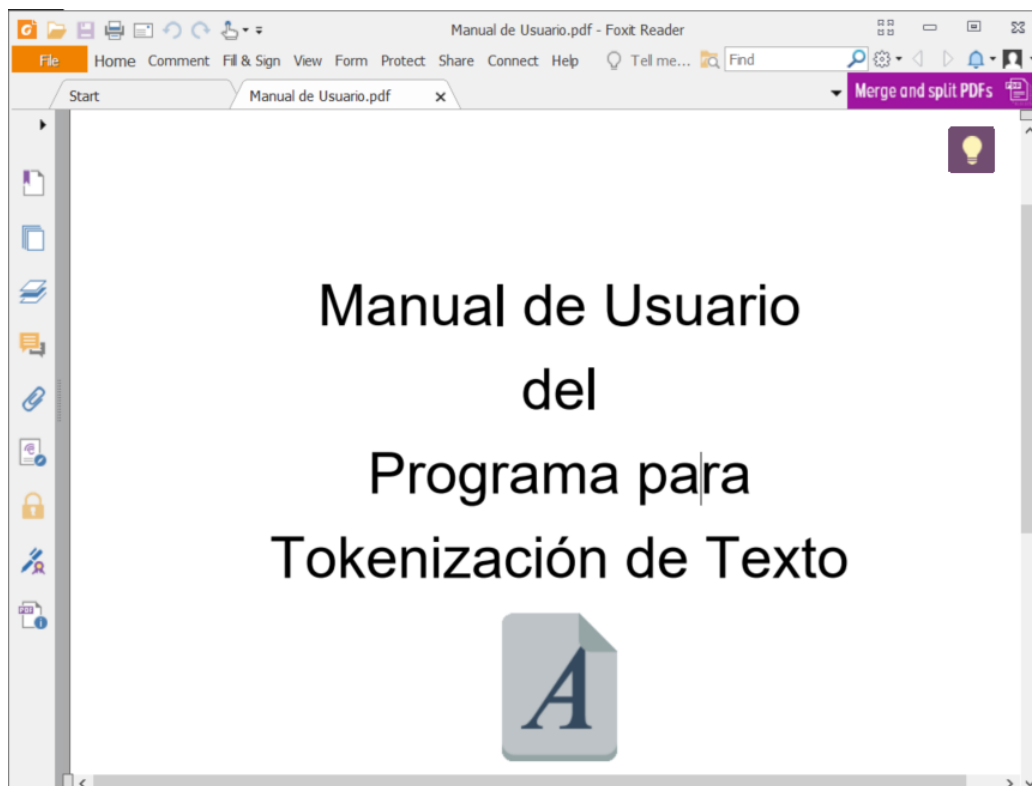


Al hacer clic sobre la opción de **“Ayuda”**, desplegarán algunas de las tareas que puede realizar el programa.

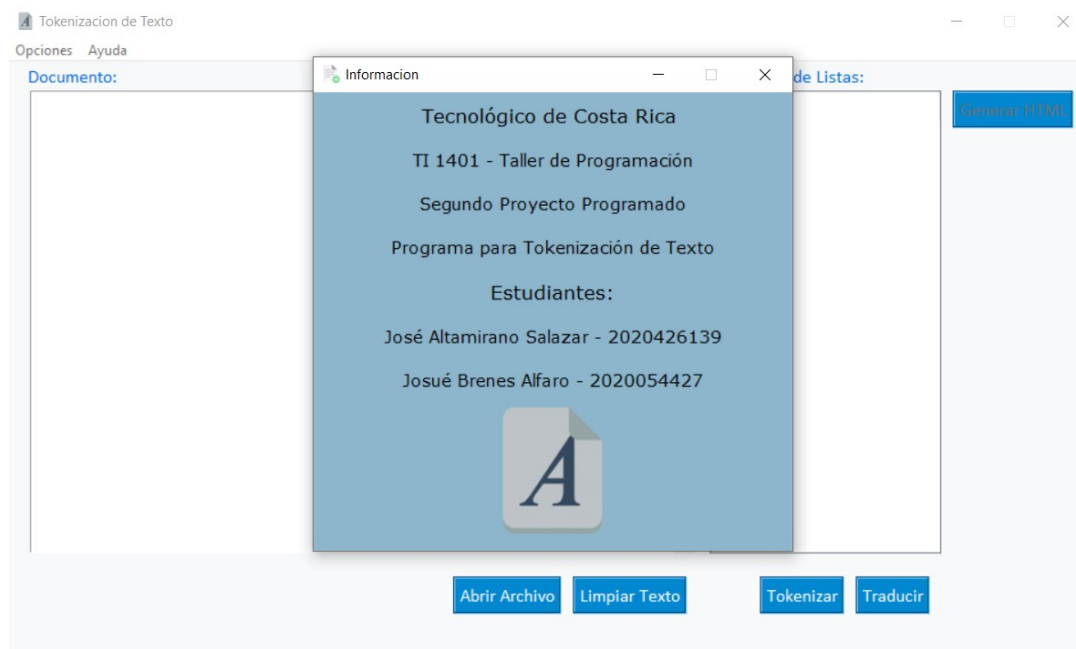


Si se selecciona la primera opción de ese submenú, **“Manual de Usuario”**, el programa le abrirá el documento que contiene el este manual de usuario. Esta opción debe ser utilizada en caso de que se requieran consultar los pasos necesarios para algún proceso o tarea del programa.





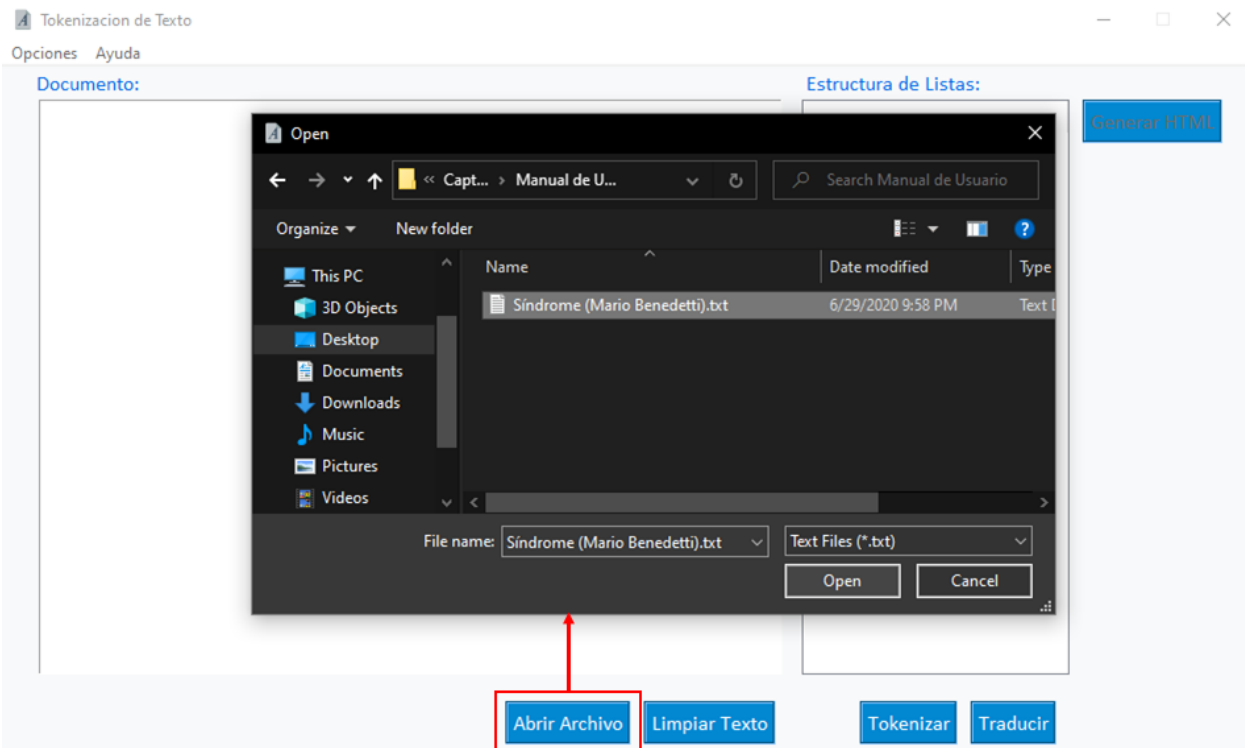
Si se selecciona la segunda opción de ese submenú, **“Acerca de”** , el programa le abrirá otra ventana que contiene información sobre el equipo que desarrolló el software.

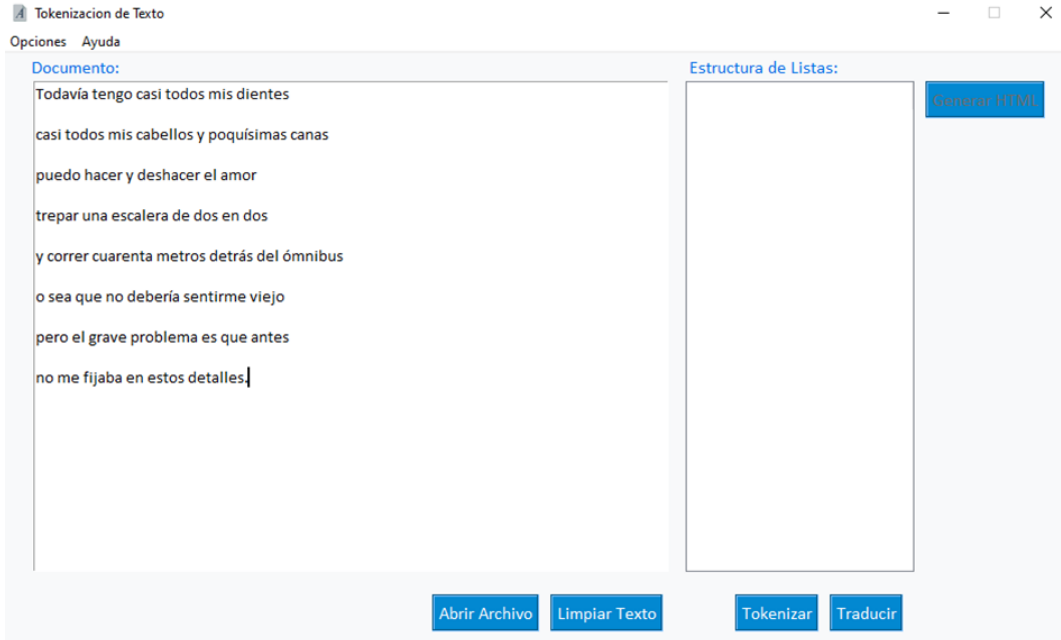


- **Tokenizar un texto**

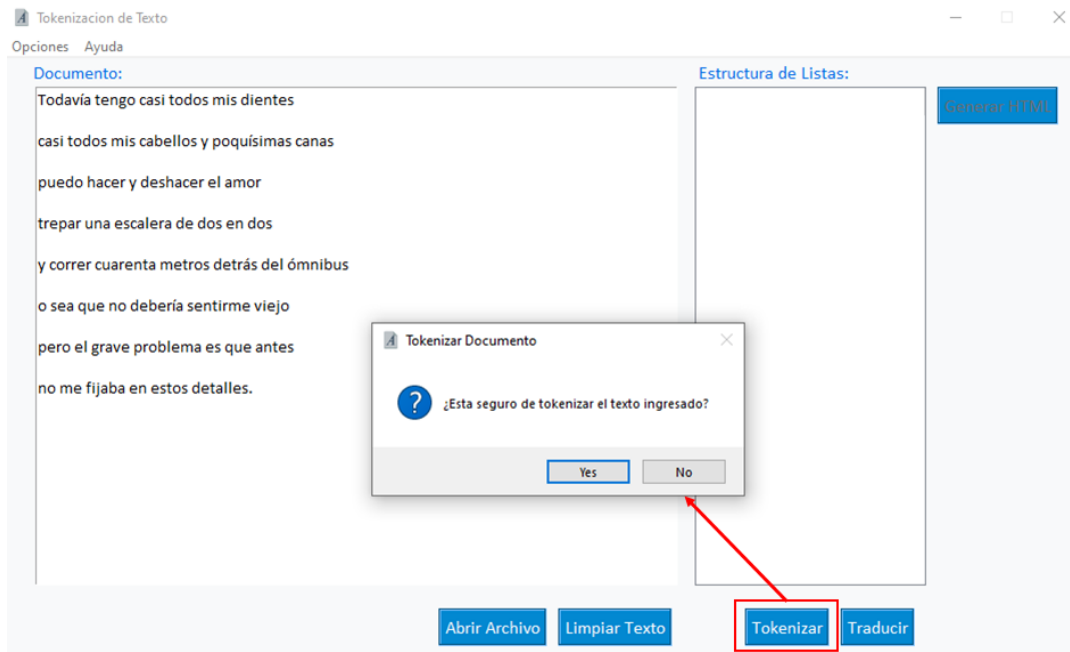
El proceso para tokenizar un texto es muy sencillo, y consta de solamente 4 pasos, los cuales se describen a continuación.

1. Ingrese texto dentro del campo dispuesto para este propósito. Para esto puede dar clic sobre el cuadro ubicado a la izquierda de la pantalla y digitarlo de forma manual o también puede seleccionar un archivo de texto. Para seleccionarlo haga clic sobre la opción del menú superior: **Opciones-->Abrir un archivo** o presione el botón **Abrir archivo** en la parte inferior de la ventana. Al seleccionar un archivo y presionar el botón “**Abrir**”, el texto contenido en este se mostrará en la ventana del programa.

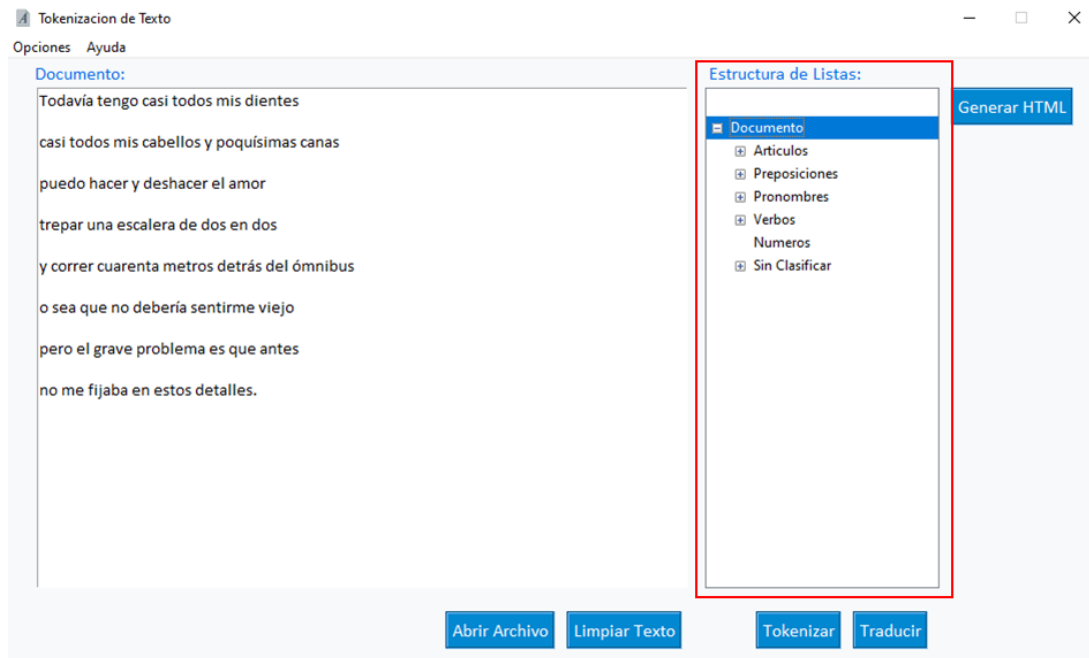




2. En caso de que requiera editar el texto extraído del archivo, puede hacerlo en este momento. Ahora bien, para poder clasificar cada una de las palabras contenidas en el texto, deberá dar clic sobre el botón “**Tokenizar**”. Al hacer esto, se le mostrará una ventana para confirmar la tarea de tokenización del texto ingresado.

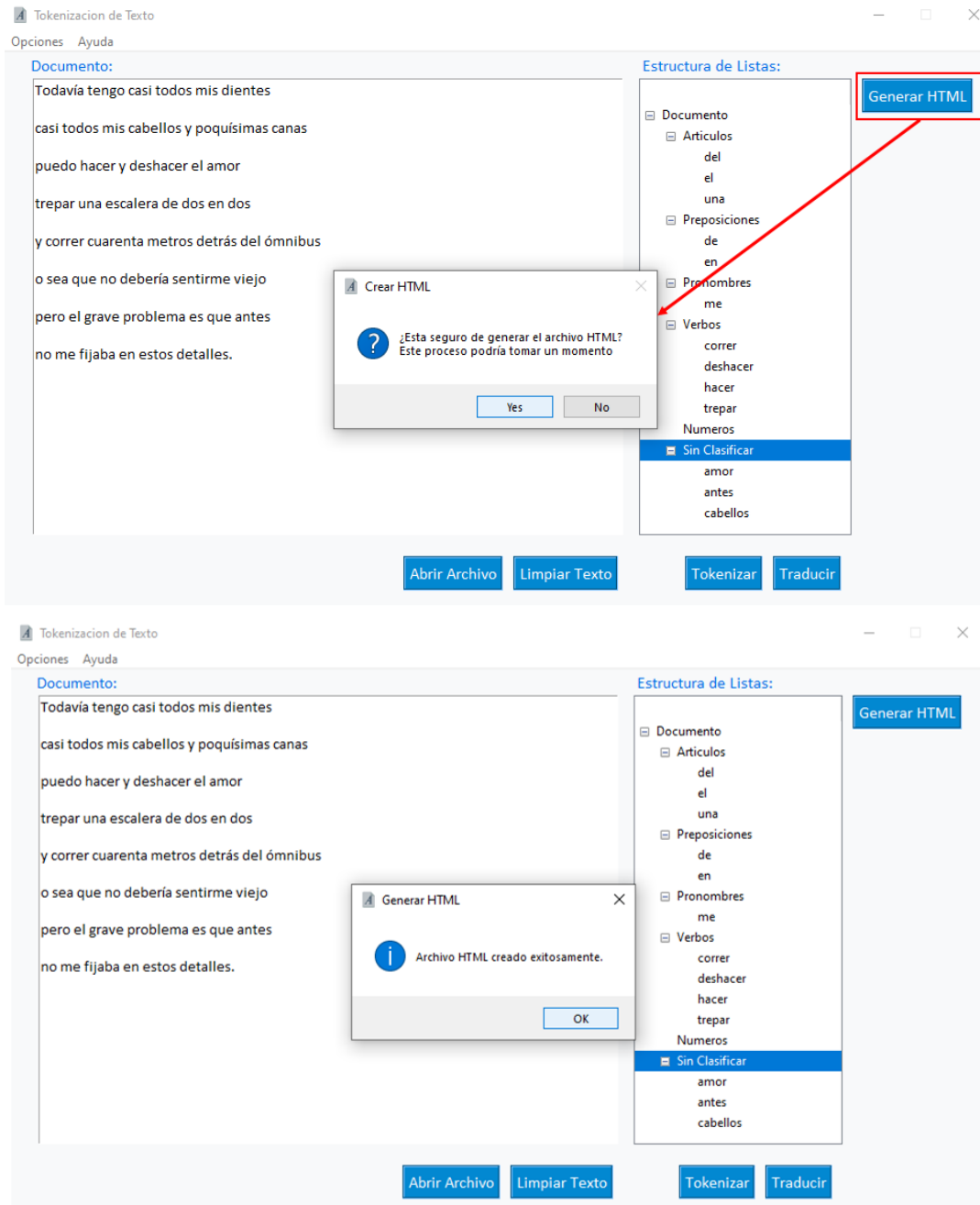


Si presiona que si desea continuar, se cargará el resultado de la operación el campo de listado de la sección izquierda de la ventana.

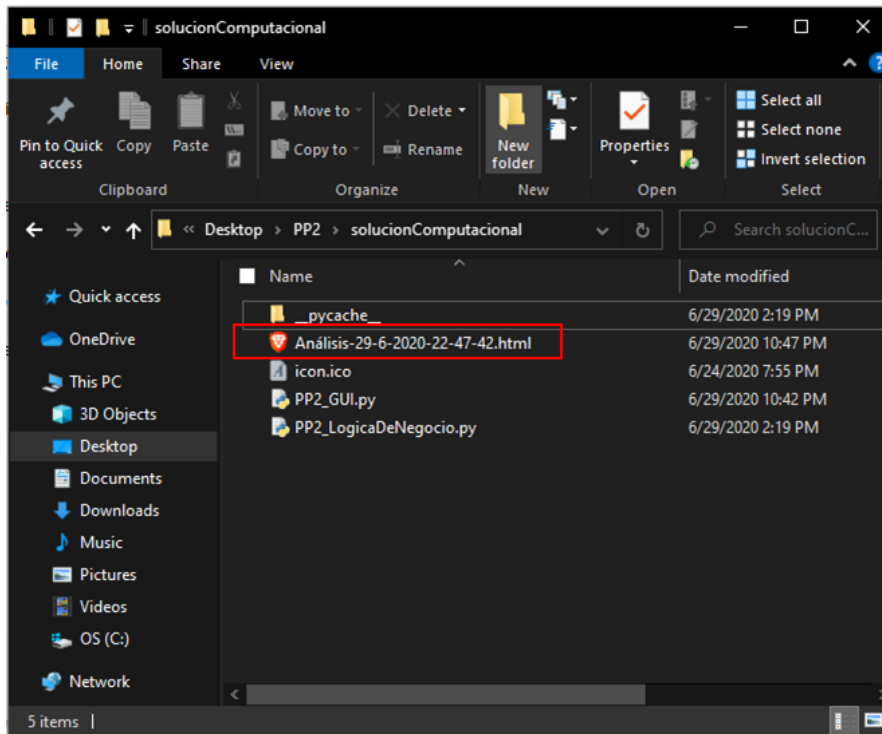


Las opciones del listado pueden ser desplegadas para visualizar todos los elementos contenidos en cada categoría

3. Finalmente, para guardar el proceso de la tokenización generado hasta este punto, podrá dar clic sobre el botón “**Generar HTML**”, el cual hasta este punto se ha mantenido deshabilitado. Al dar clic sobre este se mostrará otra ventana para confirmar que se desea generar el archivo correspondiente, y si se acepta la operación luego de unos segundos se le indicará que el archivo se ha generado de forma correcta.



4. Para poder visualizar el archivo generado, diríjase a la carpeta en donde se encuentra almacenado el archivo ejecutable del programa, y allí podrá encontrar el archivo correspondiente a clasificación recién generada., al dar clic sobre éste el mismo se abrirá en el navegador predeterminado de su equipo, y podrá visualizar la clasificación de los tokens tabulada. El archivo se encuentra guardado con un nombre que incluye la fecha y la hora en fue generado.



Clasificación de Elementos

File | C:/Users/joma2/Desktop/PP2/solucionComputacional/Análisis-29-6-2020...

Contenido Analizado

Todavía tengo casi todos mis dientes casi todos mis cabellos y pocas canas puedo hacer y deshacer el amor trepar una escalera de dos en dos y correr cuarenta metros detrás del autobús o sea que no debería sentirme viejo pero el grave problema es que antes no me fijaba en estos detalles.

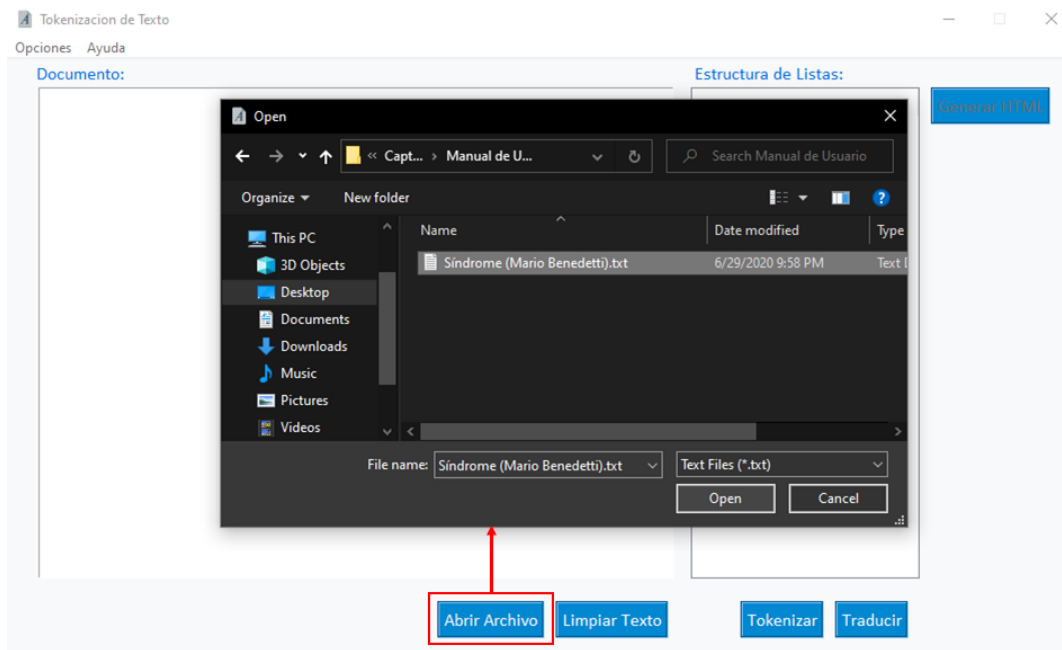
Análisis del documento

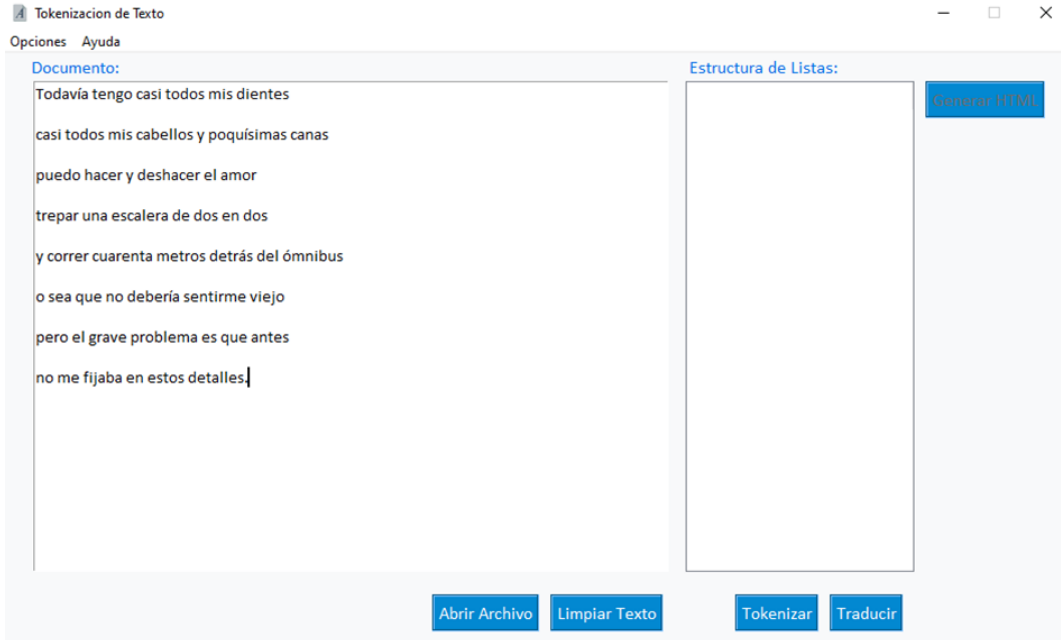
Artículos	Preposiciones	Pronombres	Verbos	Numeros	Sin Clasificar
del	de	me	correr		amor
el	en		deshacer		antes
una			hacer		cabellos

- **Visualizar la traducción de un texto**

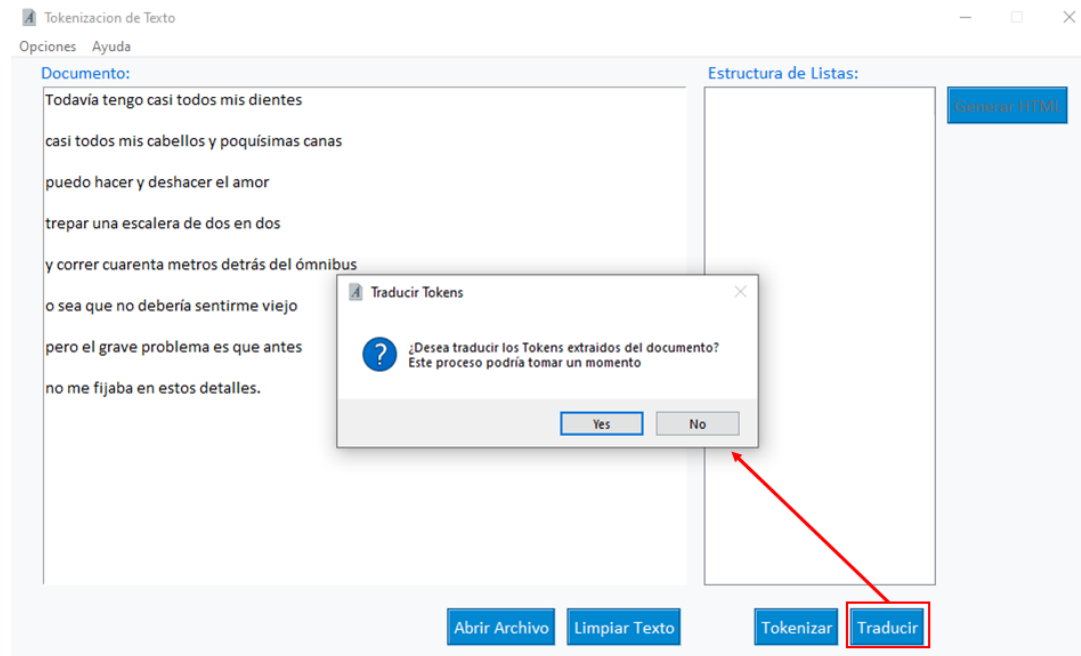
En caso de que se requiera verificar o visualizar la traducción de los tokens dentro del programa, debe seguirse un proceso similar al anterior. De igual manera, los pasos necesarios para completarlo se describen a continuación.

1. Ingrese texto dentro del campo dispuesto para este propósito. Para esto puede dar clic sobre el cuadro ubicado a la izquierda de la pantalla y digitarlo de forma manual o también puede seleccionar un archivo de texto. Para seleccionarlo haga clic sobre la opción del menú superior: **Opciones-->Abrir un archivo** o presione el botón **Abrir archivo** en la parte inferior de la ventana. Al seleccionar un archivo y presionar el botón **“Abrir”**, el texto contenido en este se mostrará en la ventana del programa.

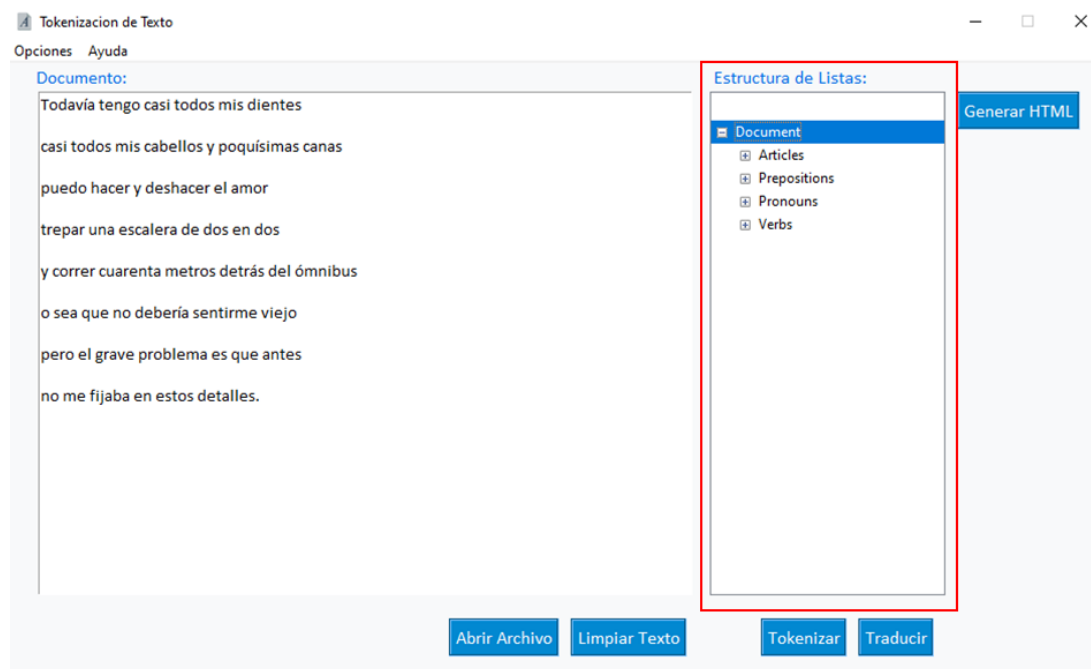




2. En caso de que requiera editar el texto extraído del archivo, puede hacerlo en este momento. Ahora bien, para poder clasificar cada una de las palabras contenidas en el texto, deberá dar clic sobre el botón **"Traducir"**. Al hacer esto, se le mostrará una ventana para confirmar la tarea de traducción del texto ingresado.



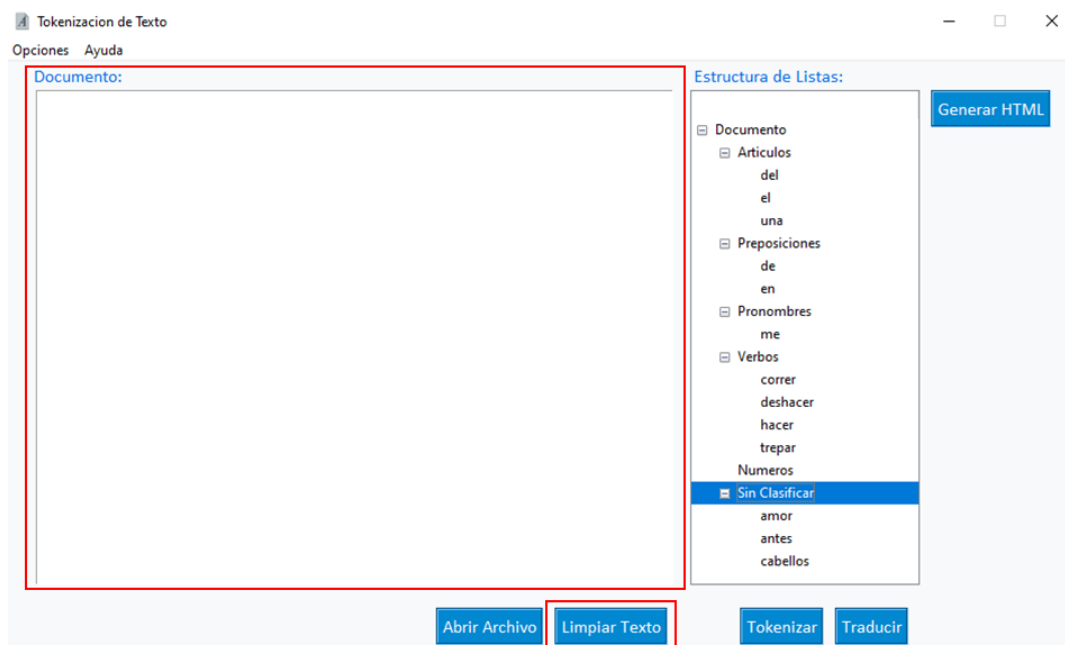
Si presiona que, si desea continuar, se cargará el resultado de la operación el campo de listado de la sección izquierda de la ventana.



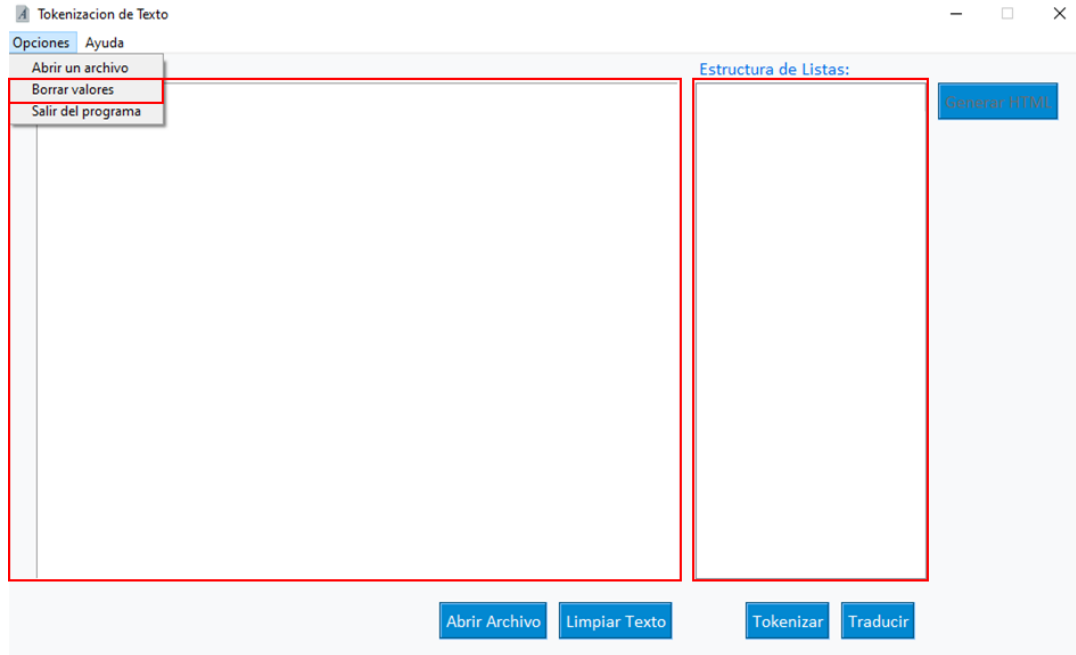
Las opciones del listado pueden ser desplegadas para visualizar todos los elementos contenidos en cada categoría.

- **Borrar los valores ingresados**

En caso de que se requiera eliminar los valores contenidos dentro del programa, se dispone de dos opciones para ejecutar los procesos relacionados a esta tarea. Si solamente es necesario eliminar el texto ingresado en la sección de documento, basta con presionar el botón de **Limpiar texto** de la parte inferior de la ventana. Tenga en cuenta que esta función solamente elimina el texto contenido dentro de la sección mencionada y no así la clasificación realizada en caso de que el texto haya sido tokenizado.

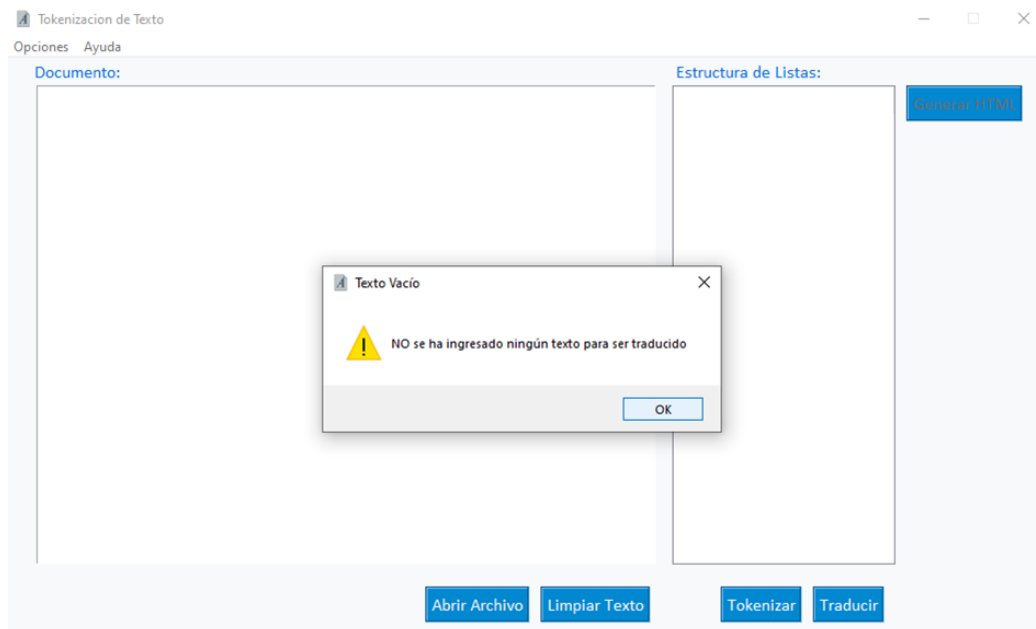
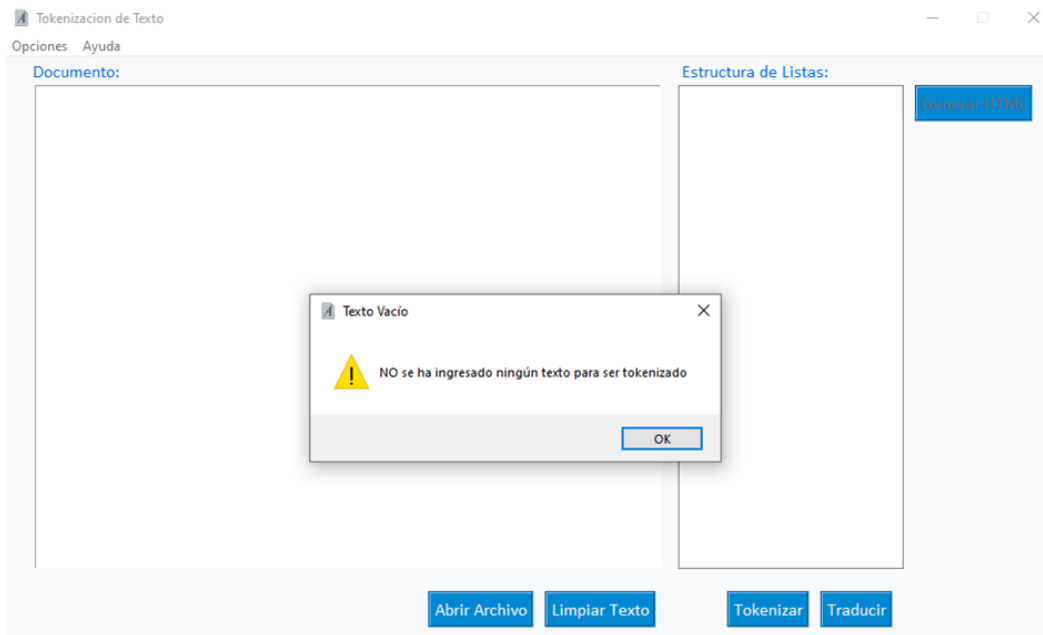


Ahora bien, en caso de que se requiera eliminar por completo todos los valores contenidos dentro del programa en la sesión activa, basta con seleccionar la opción de **Borrar valores**, disponible en el menú superior de la ventana. Esta opción eliminará tanto el texto ingresado como la clasificación generada por el programa.



- **Posibles errores dentro del programa.**

En caso de realizar una acción indebida o que al momento de solicitar la realización de alguna tarea no se hayan incluido toda la información necesaria, el programa le indicará mostrando mensaje de alerta. A continuación, se incluyen ejemplos de los mensajes desplegados por el programa en caso de que no se haya ingresado texto tanto para la tokenización de un texto como para su traducción.



En caso de que en alguno de los procesos antes mencionado, o incluso al intentar generar el archivo HTML no haya sido completado de forma correcta, el programa también le indicará al usuario cuando se produzca un error de este tipo. Ejemplo de estas salidas se incluyen las siguientes imágenes.

